Office of Science
U.S. DEPARTMENT OF ENERGY

Advanced Scientific Computing Research

FY2005 Accomplishments

Network Research

TeraPaths: A QoS Enabled Collaborative Data Sharing Infrastructure for Peta-scale Computing Research

PIs: Bruce Gibbard, Dantong Yu*, Brookhaven National Laboratory
Senior Personnel: Shawn Mckee, University of Michigan

Summary

A DOE MICS/SciDac funded project, TeraPaths, along with its managed network resources, has given rise to data-on-demand capability from the scientific apparatus and data repositories to computing centers used for data analysis. In particular, this project deployed and prototyped the use of differentiated networking services based on a range of new transfer protocols in support of global data movement. The managed network capability being enabled by this project will be integrated, and scheduled as part of Grid computing systems, along with the managed CPU and storage resources, to enhance the overall performance and efficiency of DOE computing facilities. We demonstrated TeraPaths' effectiveness in data transfer activities in Brookhaven National Lab.

Project Objectivities

The primary goal of this project is to investigate the use of advance network technologies, such as Local Area Network (LAN) Quality of Service (QoS) and ***M****ulti****p****rotocol* ***L****abel* ***S****witching* (MPLS), in high energy physics data intensive distributed computing environment. A number of corollary objectives will also be achieved.

1) Expertise in MPLS based QoS technology will be developed which will be important to ATLAS (one of four experiments at Large Hadron Collider) and the high-energy physics community more generally.
2) At Brookhaven National Laboratory the ability to dedicate an equitable fraction of the available network bandwidth via MPLS/LAN QoS to ATLAS Tier 1 data movement will assure adequate throughput and limit its disruptive impact on BNL's heavy ion physics experiment funded by DOE nuclear physics program and other more general laboratory network needs.
3) The project will enhance technical contact between the ATLAS Tier 1 at BNL and its network partners including the Tier 0 center at CERN (EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH), ATLAS Tier 2's and other members of the Open Science Grid community of which it is a part.

Project Accomplishments

TeraPaths project is enabling data transfers with guarantees of speed and reliability that are crucial to applications with deadlines, expectations and critical decision-making requirements. Brookhaven National Lab needs to do RHIC (Relativistic Heavy Ion Collider) production data transfers and LHC (Large Hardon Collider) Monte Carlo

*Bruce Gibbard (631)-344-7969, gibbard@bnl.gov, Dantong Y: (631)-344-3042, dtyu@bnl.gov
Project WebSite: http://www.atlasgrid.bnl.gov/terapaths

challenges between BNL and the remote collaborators. The aggregate of their peak network requirements is beyond BNL capacity. Our project can modulate LHC data transfers to opportunistically utilize available bandwidth, ensuring that the RHIC production data transfer is not impacted.

During RHIC 2005, about 270 Tera-byte of data (3.5 billion proton-proton events) were directly moved from the on-line system to Japan over a period of 11 weeks. All raw event data can be delivered to collaborations a few hours after their generation, regardless of the collaborators' physical distances to BNL, local at BNL or oversea, providing equal opportunities for all participants.

In FY 2005, we acquired capability to configure network equipment to dedicate fractions of available bandwidth via QoS for various BNL data movement/replications and limit their disruptive impact upon each other. We implemented non-intrusive QoS mechanisms to enhance overall network performance and verified their effectiveness via experiments.

Software development proceeded smoothly. The software could automate the QoS configuration in network paths and negotiate network bandwidth with the remote network resource manger on behalf of end users. Our system architecture could be easily integrated with other network management tools to provide a complete end-to-end QoS.

We have been closely collaborating with other MICS/SciDAC funded network projects to leverage software development and integrate our system with theirs to provide end users managed network services.

The Impact to Specific DOE Science Application:

Our project is a central component of a responsive, managed infrastructure for eScience. Such an infrastructure can ensure that physics event data can be delivered to collaborators a few hours after its generation, regardless of the collaborators' physical distances to data source, local or oversea, providing equal opportunities for all participants.

Future Work

Our future work will concentrate on strategically scheduling network resource to shorten the transfer time for mission critical data relocation, thus reducing the error rates which are proportional to the transfer time. We will thereby enhance the effectiveness of network utilization. We will manage network resources which typically span many administrative domains, a unique characteristic compared with CPU and storage resource. The overall goal remains providing a robust, effective infrastructure for High Energy and Nuclear Physics.

For further information on this subject contact:

Dr. Thomas Ndousse, Program Manager
Mathematical, Information, and
Computational Sciences Division
Office of Advanced Scientific Computing Research
Phone: 301-555-3691
tndousse@er.doe.gov

Advanced Scientific Computing Research
FY 2005 Accomplishment

Real Science – *"Security and Policy for Group Collaboration"*

Ian Foster* Frank Siebenlist, Rachana Ananthakrishnan
Mathematics and Computer Science Division, Argonne National Laboratory

Summary

The "Security and Policy for Group Collaboration" project develops Grid Security Infrastructure, i.e., software and tools to enforce the required security policies in DOE's major distributed science projects. Thousands of scientists use these tools to achieve secure access to remote data and computational services.

Today's science and engineering are distributed: the data, computers, software, instruments, and colleagues needed to solve challenging problems are located across the country. Fortunately, high-speed networks provide for remote access—*but only if security problems are overcome*. Specifically, both facility operators and scientists need authentication mechanisms (to establish the identity of remote users) and authorization mechanisms (to apply the policies concerning who is allowed to access specific resources). These mechanisms must be secure against a wide variety of attacks.

DOE support has enabled major progress on this problem. Specifically, the **Security and Policy for Group Collaboration** project has produced authentication and authorization algorithms and software that have been adopted by dozens of major distributed science projects. These projects have in turn profited from the availability of high-quality secure authentication and authorization mechanisms to achieve significant advances in distributed science. We give three examples of such projects here.

The **DOE Earth System Grid** data portal (Figure 1) has used Grid Security Infrastructure (GSI) mechanisms to register over one thousand climate researchers as users during the past year. These users have downloaded tens of terabytes of data from Earth System Grid sites and produced 250 publications from International Panel on Climate Change data this year alone.

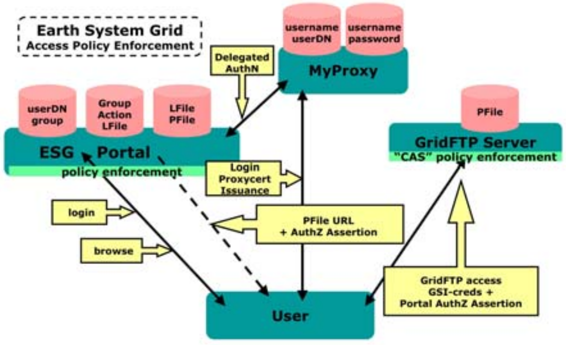

Figure 1. Earth System Grid Security

Participants in the **DOE Particle Physics Data Grid** collaboration have used GSI mechanisms to enable not only high energy and nuclear physicists but also biologists and chemists to harness computers and storage at 50 sites across the U.S. for large-scale distributed data analysis. This work

* 630-252-4619, itf@mcs.anl.gov

has reduced significantly the time required to analyze data from physics and biology experiments.

The **DOE Fusion Collaboratory** (Figure 2) has used GSI mechanisms to enable secure remote access to advanced fusion codes. The resulting deployment of advanced simulation codes as services has enabled in a dramatic increase in the use made of advanced simulation capabilities, increasing by an order of magnitude the number of simulations performed over the past year. They have also created a secure remote control roolm

Each of these three projects has reported significant achievements of their own. These achievements were all made possible by the availability of advanced security capabilities provided by the **Security and Policy for Group Collaboration** project.

For further information on this subject contact:
Dr. Thomas Ndousse, Program Manager
Mathematical, Information, and
Computational Sciences Division
Office of Advanced Scientific Computing Research
Phone: 301-555-3691
tndousse@er.doe.gov

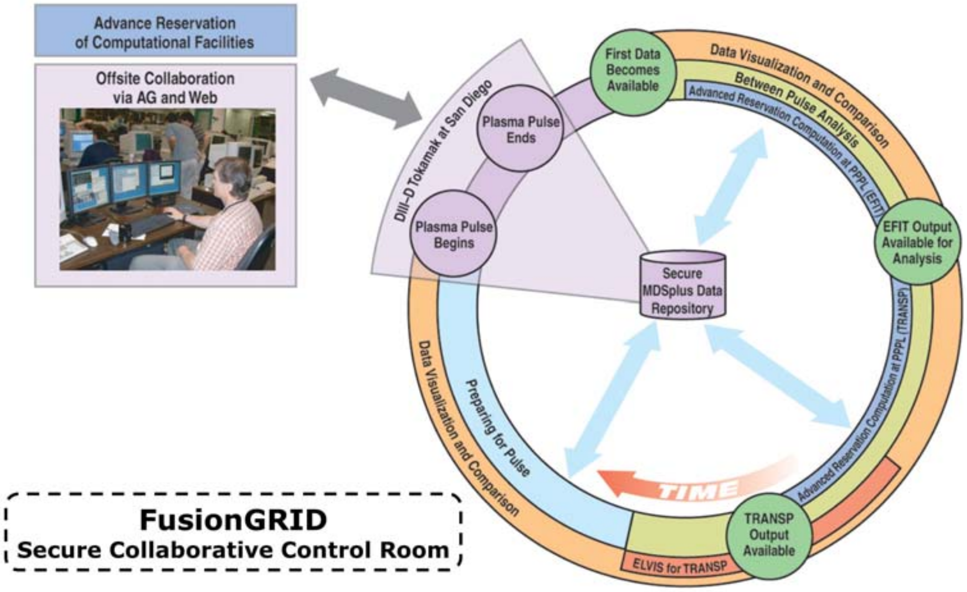
Advance Reservation of Computational Facilities
Offsite Collaboration via AG and Web
DIII–D Tokamak at San Diego
Plasma Pulse Ends
Plasma Pulse Begins
First Data Becomes Available
Data Visualization and Comparison
Between Pulse Analysis
Advanced Reservation Computation at PPPL (EFIT)
EFIT Output Available for Analysis
Secure MDSplus Data Repository
Advanced Reservation Computation at PPPL (TRANSP)
TRANSP Output Available
ELVIS for TRANSP
TIME
Preparing for Pulse
Data Visualization and Comparison
FusionGRID
Secure Collaborative Control Room

Figure 2. FusionGRID Remote Control Room

Lambda Station

Don Petravick[*], PI, Fermi National Accelerator Laboratory
Harvey Newman, co-PI, California Institute of Technology
Phil Demar, Senior Researcher, Fermi National Accelerator Laboratory
Matt Crawford, Project Manager, Fermi National Accelerator Laboratory

Summary

Lambda Station is a network path selection and setup service that routes selected traffic between major computing resources over available high-impact wide area networks. In so doing, it communicates with a Lambda Station at the intended peer site, performs any needed reservation function on the wide area network, and the two Lambda Stations may alter their sites' internal routing and border router access rules to link the major applications across the high performance network. During fiscal year 2005 all these functions were integrated into a working prototype, which has been exercised over UltraScience Net and other networks.

High-Energy Physics, and other fields of science important to the Department of Energy, are supported by very large-scale computational facilities processing, storing and moving petabyte-scale data sets. Special high performance networks are available to science programs, but not universally available for all pairs of communicating endpoints, or at all times. Dedicating the computational assets to the "high-impact" networks is seldom feasible, since the same assets serve small many consumers over general-purpose networks as well as large stakeholders. Dual-homing the assets to production and high-impact networks can be prohibitively expensive, introduces difficult operational complexities, and still does not provide the path robustness that might be expected from an investment in dual-homing.

The Lambda Station solution exploits the generally untapped "policy routing" capabilities of the site network infrastructure to route selected flows to the high-impact network while leaving other traffic's routing unaffected. This re-routing can be instituted even while the traffic is flowing and undone whenever necessary – if, for example, the high-impact path fails or the reserved time period elapses.

An application requests Lambda Station's services by providing a specification of the traffic that is to receive special handling. (This includes a designation of the endpoint resources, and possibly protocol and port information.) All the knowledge of intra- and inter-site topology is encapsulated in Lambda Station's configuration and access to the high impact network can be controlled to a granularity of individual flows. By dynamically configuring site routing at this level, advanced research networks are made available to production applications rather than just to demonstrations.

The immediate application – and the development and testing environment – is LHC physics data movement. Although a dedicated network has been acquired for the initial transfer of data from the Tier-0 site at CERN to the US Tier-1 centers (FNAL and BNL), LHC-dedicated network paths to

[*] (630) 840-3935, petravick@fnal.gov

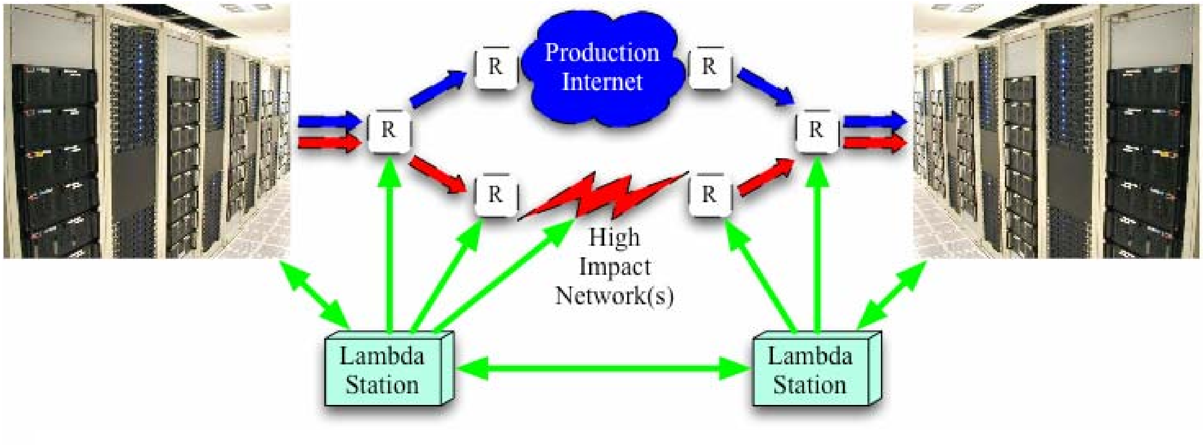

Figure 1. A Lambda Station at each site adjusts internal routing and border packet filtering rules. The Lambda Station on the initiating side is responsible for setup of the high-impact WAN path.

Tier-2 centers do not generally exist, or are time-shared among different phases of the data analysis cycle. In rare cases, a Tier-2 site may have access to multiple high-performance wide-area networks, perhaps subject to reservation. The network bandwidth of a Tier-3 site is even more constrained and such a site requiring a large data transfer may need to negotiate use of a network path. Lambda Station addresses all of these issues.

Fiscal year 2005 was the first year of the Lambda station project. During this year we have created a fully functional prototype of Lambda Station and implemented the client calls in multiple programming languages. The prototype is capable of configuring Cisco site and border routers to access a reserved path over UltraScience Net. Diagnostic applications lstraceroute and lsiperf have been created, which are versions of traceroute and iperf that make calls to Lambda Station. Lambda Station calls have been added to the gridftp client and server which are integral to the reference implementation of SRM, the Storage Resource Manager, which is the data movement control protocol already in use at FNAL and all US CMS Tier-2 sites and at BNL and most US ATLAS Tier-2 sites.

Lambda Station presents a Service-Oriented (SOAP) interface to clients and to peer Lambda Stations, making it amenable to integration into larger workflows. Our work to date has been well received in the advanced optical networking community and we look forward to working with the UCLP (User-Controlled Light Path) and GLIF (Global Lambda Integrated Facility) researchers.

During the coming year we will integrate the Lambda Station client functions into dCache, the hierarchical storage manager underlying the US LHC storage systems. We will add support for dynamic configuration of Force10 routers and perhaps another brand. We will port all modules to Java (some are currently in Perl) and produce a readily installable, documented software kit for installation at other sites.

For further information on this subject contact:
Dr. Thomas Ndousse, Program Manager
Mathematical, Information, and Computational Sciences Division
Office of Advanced Scientific Computing Research
Phone: 301-555-3691
tndousse@er.doe.gov

Advanced Scientific Computing Research
FY 2005 Accomplishment

UltraScience Net – *"Networking for Expediting Scientific Discovery"*

Nageswara S. V. Rao [*], William R. Wing, Oak Ridge National Laboratory

Summary

The next generation DOE large-scale science projects require unprecedented network capabilities for petabyte data transfers, and effective collaborations, visualizations, computational steering and instrument control across the country. DOE UltraScienceNet is a powerful and unique experimental network testbed that enables the development of these advanced networking and related application technologies at the scale and scope needed for these large-scale applications. This testbed provides on-demand and in-advance scheduled dedicated high bandwidth channels for large data transfers, and also high resolution and stable channels for finer control operations. The backbone consists of dual 10Gbps links from Atlanta to Chicago to Seattle to Sunnyvale. Its secure control plane enables the scheduling of dedicated bandwidth channels, and authenticated and encrypted signaling to network devices. It currently provides connectivity to FNAL, PNNL, ORNL, SLAC, CalTech and ESnet. This network has been completely deployed, and has been successfully utilized for high bandwidth data transfers and dedicated customized channels to ORNL Cray X1 supercomputer.

The next generation of 100-1000 Teraflop supercomputers needed for DOE large-scale science computations hold an enormous promise for large-scale computational science projects from fields as diverse as earth science, high energy and nuclear physics, astrophysics, fusion energy science, molecular dynamics, nanoscale materials science, and genomics. These computations are expected to generate hundreds of petabytes of data, which must be transferred, visualized and steered by geographically distributed teams of scientists. In the experimental sciences area, DOE operates several extremely valuable experimental facilities, for example, the Spallation Neutron Source. The ability to remotely conduct experiments, then transfer large measurement data sets, can significantly enhance the productivity of scientists and facilities. Indeed, the above networking capabilities add a whole new dimension to the access of these computers and user facilities, by eliminating the "single location, single time zone" bottlenecks.

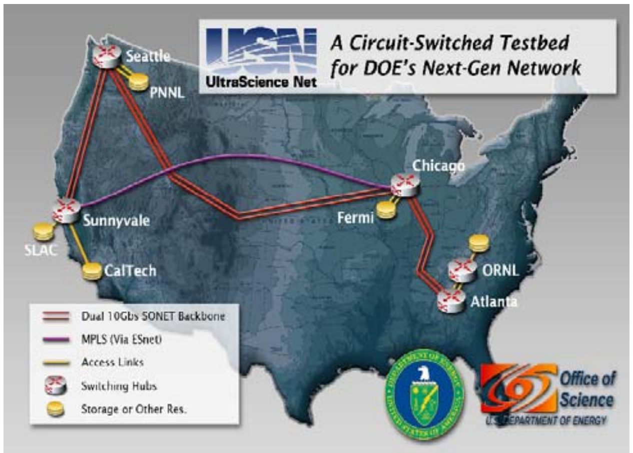

Figure 1. USN's dual 10bps links span Atlanta, Chicago, Seattle and Sunnyvale. It connects ORNL, PNNL, FNL, SLAC, ESnet and Caltech.

The required network capabilities in terms of both usable high bandwidth and precision control are extremely difficult to support over the current ESnet and Internet. The existing testbeds and networks for providing such channels typically have a very small footprint or bandwidth, offer only a single lambda or sub-lambda, and are not field hardened and optimized for wide-area deployments and cyber defense.

[*] Telephone Number: (865) 574-7517, E-mail address: raons@ornl.gov

UltraScienceNet is an experimental network research testbed dedicated to the development of networking capabilities needed for the above DOE large-scale science applications. USN backbone consists of dual SONET OC192 10Gbps links that span Atlanta, Chicago, Seattle and Sunnyvale as shown in Figure 1. It currently provides connectivity to FNAL, ORNL, PNNL, SLAC, ESnet and CalTech. It provides on-demand dedicated channels to connect resources at multi-, single- and sub-lambda resolution. Users can utilize hosts located at UltraScienceNet edges or connect to it through their own specialized connections. Various types of protocols, middleware and application research projects in support of DOE large-scale science applications can make use of the provisioned dedicated circuits. Through these activities, USN serves as a technology pathfinder and developmental platform for advanced networking technologies, which will eventually be deployed in production ESnet upon maturation.

USN control-plane and management-plane are deployed using a Virtual Private Network that provides protection against cyber attacks. The control-plane is implemented using a centralized scheduler that (a) maintains the state of bandwidth allocations on each link; (b) accepts and grants requests for current and future channels to applications; and (b) sends signaling messages to switches as per the schedule for setting up and tearing down the dedicated channels.

A full deployment of USN backbone was completed in August 2005. In the current configuration, it provides on-demand end-to-end guaranteed circuits with capacities ranging from 50 Mbps to 20 Gbps within minutes of setup times. Such capability is in stark contrast with the Internet where the shared connections are statically provisioned, and as a result, the bandwidth is neither guaranteed nor stable

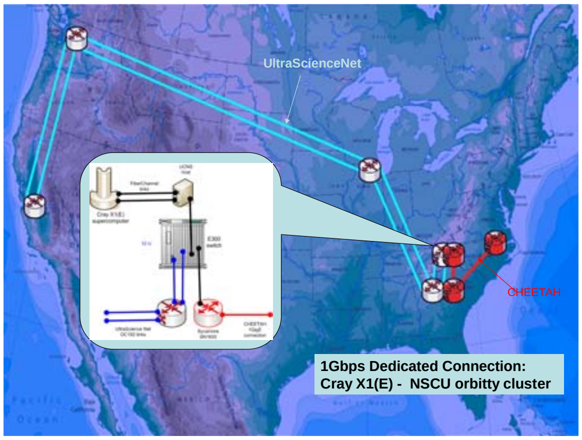

Figure 2. USN provides high-performance dedicated connection to ORNL Cray X1 supercomputer.

During the past year, USN has enabled researchers at FNAL to develop and test advanced data transfer technologies, which achieved data rates 15 times faster compared to their production ESnet connection. In collaboration with NSF CHEETAH network, USN provided a dedicated channel to ORNL Cray X1 supercomputer as shown in Figure 2, and achieved speeds of 1.4 Gbps, about 28 times faster than default rate.

USN has attracted considerable attention from networking and application community, and papers describing USN technologies have been presented at conferences and published in leading journals. Its architectural design has been adopted by the upcoming LHCnet. Its design and control-plane technologies are under consideration for adoption by NASA and DoD.

For further information on this subject contact:
Dr. Thomas N. Ndousse, Program Manager
High-Performance Networking Program
Mathematical, Information, and Computational
Science Division
Office of Advanced Scientific Computing Research
Phone: 301-903-9960; email: tndousse@er.doe.gov

Advanced Scientific Computing Research
FY 2005 Accomplishment

UltraScienceNet Genomics Testbed

Thomas P. McKenna *, Pacific Northwest National Laboratory

Summary

A network research testbed is being designed and implemented to enable computational genomics applications in support of the Department of Energy (DOE) Genomes for Energy and Environment research program. The main goal of the testbed is to develop and deploy networking technologies needed for remote instrument control and real-time streaming of large-scale data for genomics applications operating in the framework of DOE's distributed genomics facilities. Researchers and developers will be provided with an experimental infrastructure for on-demand, dedicated bandwidth to genomics applications. This work will facilitate genomic discovery by enabling scientists throughout the world to research, teach, and collaborate, utilizing state-of-the-art instruments.

Development and deployment of networking technologies will be accomplished through basic real-time control protocol research, application of other research efforts in visualization and data transport, development of an API for scientific applications to interface to different transport protocols, and prototype implementation and testing using dedicated channel capabilities of the UltraScienceNet. The testbed application encompasses the remote operation of a multi-spectral confocal microscope, streaming high-quality image data for storage and processing, and near-real-time visualization of experimental results to allow immediate decisions about the future course of the experiment. The instrument enables the examination of cell-level functions and structures, which are critically important for pursuing a systems approach to molecular and cellular biology. The genomics testbed is being implemented over the UltraScienceNet initially connecting Pacific Northwest National Laboratory (PNNL) and Oak Ridge National Laboratory (ORNL) with the ability to scale up to include other DOE genomics research facilities in the future.

In recent months, several key milestones have been reached by the UltraScienceNet (USN) Testbed project in the areas of application development, connectivity, and software configuration management.

USN Application Development Milestones

The proof-of-concept is being developed in a three-phase process, parts of which can be performed concurrently, that can be leveraged to build a demo. Phase one entails establishment of the connectivity of the CaMatic software from a remote workstation to a server running software that emulates the confocal microscope so that progress can be made independent of the scheduled usage of the actual microscope, and development software

* 509-372-6180, thomas.mckenna@pnl.gov

can be tested without risk of damage to the actual microscope hardware.

This new remote CaMatic software in concert with the microscope emulator module tests our ability to control the microscope functionality via the . The first complete version of USNAPI software (the network layer) has been completed and is currently being tested. We have also completed building the remote version of the CaMatic software; begun testing it utilizing the UltraScienceNet Application Programmable Interface (USNAPI); and successfully tested several functions of the microscope across this network layer. We are currently testing the complete functionality of the microscope control system, and we will soon be connecting the development machines across a 1GB packet-switched network to enable a more accurate testing environment.

Phase-two of the proof-of-concept has begun as well, and includes streaming the real-time visualization imagery of the microscope via the USNAPI from the microscope to the remote workstation. We are modifying the current capture-and-display paradigm of the CaMatic software to fit a networked model.

Phase-Three of the proof of concept will begin shortly, and will prove our ability to move large amounts of data off and on the machines utilizing 10GB network cards at speeds that make good use of the UltraScienceNet bandwidth.

USN Connectivity Milestones

The connectivity team has made excellent progress getting USN connectivity up and running. We have successfully completed loopback testing between systems on the PNNL USN Force10 switch, and systems connect at ORNL to their USN Force10 switch has been completed. We have achieved USN connectivity between the Intrinsically Secure Computer Lab (the new and improved CIPAL) and the ISB2 building. Fiber between the ISB2 building, the 331 building, and the confocal microscope are now lit up. We have space planned and dark fiber available for a USN connection from the APEL building that will be occupied by the Cyber Security group.

Software Configuration Management Milestones

We have stood up a development server in the ISB2 building for software configuration management and taken a commercial approach to managing the software for this project. This includes a source code management system (Microsoft Visual SourceSafe), integrated with a bug tracking system (Seapine Software TestTrack Pro running on top of SQL Server). This system is also integrated with MS Outlook, and also allows anyone with web access to log a bug. This approach will serve us well as we scale this into a production capability and build the team.

For further information on this subject contact:
Thomas Ndousse-Fetter, Program Manager
Mathematical, Information, and Computational Sciences Division
Office of Advanced Scientific Computing Research
Phone: 301-903-9960
tndousse@er.doe.gov